Zootaxa 4706 (3): 401–407
https://www.mapress.com/j/zt/

Article

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

https://doi.org/10.11646/zootaxa.4706.3.1
http://zoobank.org/urn:lsid:zoobank.org:pub:E7F337CF-0798-451A-8651-1CFBFF2EDB7D

The disconnect between DNA and species names: lessons from reptile species in the NCBI taxonomy database

AKHIL GARG[1], DETLEF LEIPE[2,*] & PETER UETZ[1,*]
[1]*Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia 23284, USA. E-mail: peter@uetz.us*
[2]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. E-mail: leipe@ncbi.nlm.nih.gov*
**Corresponding authors*

Abstract

We compared the species names in the Reptile Database, a dedicated taxonomy database, with those in the NCBI taxonomy database, which provides the taxonomic backbone for the GenBank sequence database. About 67% of the known ~11,000 reptile species are represented with at least one DNA sequence and a binary species name in GenBank. However, a common problem arises through the submission of preliminary species names (such as "*Pelomedusa* sp. A CK-2014") to GenBank and thus the NCBI taxonomy. These names cannot be assigned to any accepted species names and thus create a disconnect between DNA sequences and species. While these names of unknown taxonomic meaning sometimes get updated, often they remain in GenBank which now contains sequences from ~1,300 such "putative" reptile species tagged by informal names (~15% of its reptile names). We estimate that NCBI/GenBank probably contain tens of thousands of such "disconnected" entries. We encourage sequence submitters to update informal species names after they have been published, otherwise the disconnect will cause increasing confusion and possibly misleading taxonomic conclusions.

Key words: Genbank, Reptile Database, DNA sequence, nomenclature, type specimens, synonymy

Introduction

Systematic biology has primarily depended on morphological characters for most of the last 250 years until DNA sequencing became widely available in the 1990s. The value of sequences was quickly realized and they have become indispensable in taxonomy since then. However, the efficient integration of taxonomic and other data types requires various databases to coordinate their data so that they can be exchanged and compared. Taxonomic and molecular databases play a critical role in this interdisciplinary effort as they provide the names and genetic basis for phylogenetic classification. In concert with other data providers, such as collection databases, they form a (more or less) well-integrated infrastructure for the life sciences (**Fig. 1**).

In order to improve data exchange between these data providers we investigated the relationship between two critical resources, namely a taxonomic database, represented by the Reptile Database (RDB, Uetz *et al*. 2019), and the NCBI taxonomy database ("NCBI TaxonDB"), which provides the taxonomic backbone for the DNA sequence database GenBank. GenBank and the NCBI taxonomy database are operated by the National Center for Biotechnology Information, a unit of the National Institutes of Health in the United States (Federhen 2012; Sayers *et al*. 2019). While GenBank obtains its data directly from data submitters, the NCBI taxonomy database extracts its data from GenBank and updates its taxonomy using the primary literature and outside databases such as the Reptile Database; taxonomic updates are then fed back into GenBank which, for instance, updates taxonomic names in GenBank sequence records. The Reptile Database, by contrast, obtains its data from the primary and secondary literature by manual curation, author submissions, and to some extent by importing data from other databases (e.g. taxonIDs from the NCBI taxonomy database).

There are obviously differences in scope between the two databases. For example, NCBI lists the names of hybrids and the names of extinct taxa. For instance, *Cylindraspis indica* (Schneider, 1783), an extinct giant tortoise,

Accepted by A. Bauer: 21 Oct. 2019; published: 10 Dec. 2019

is found in GenBank because there is sequence data available but extinct taxa are usually not covered by the Reptile Database.

Besides such differences in scope, our analysis revealed significant discrepancies but also strategies to find and resolve mismatches among species names. In this study we have focused on the problem of how species names, specifically reptile names (from the Reptile Database), are mapped to names in the NCBI taxonomy. We looked at datasets from 2016 (Uetz & Garg 2017), mid-2018 and again in July 2019 to see whether and how data changes. In fact, we have manually cleaned up both datasets (in the Reptile Database and NCBI taxonomy) to make them more consistent, but numerous conflicts remain whose resolution will require the contribution of authors and other people who submit data to GenBank ("data submitters"). This analysis revealed a particular weakness of sequence data, which derives from the fact that many taxonomic studies produce DNA sequences without knowing the identity or status of the species under investigation. Given the central role of DNA sequences for systematic biology, our study should be relevant to many other areas, which use such data directly or indirectly, such as phylogenetics, conservation and environmental policy. For instance, if a species is split into two or more species, based on DNA sequence data, this will not only affect the taxonomy, but also policies to protect these species, and collection managers or ecologists studying these species (Fig. 1).

We believe that our findings can be extrapolated to most taxonomic databases and their relationships to both the NCBI taxonomy and GenBank, and likely also to other taxonomic data resources.

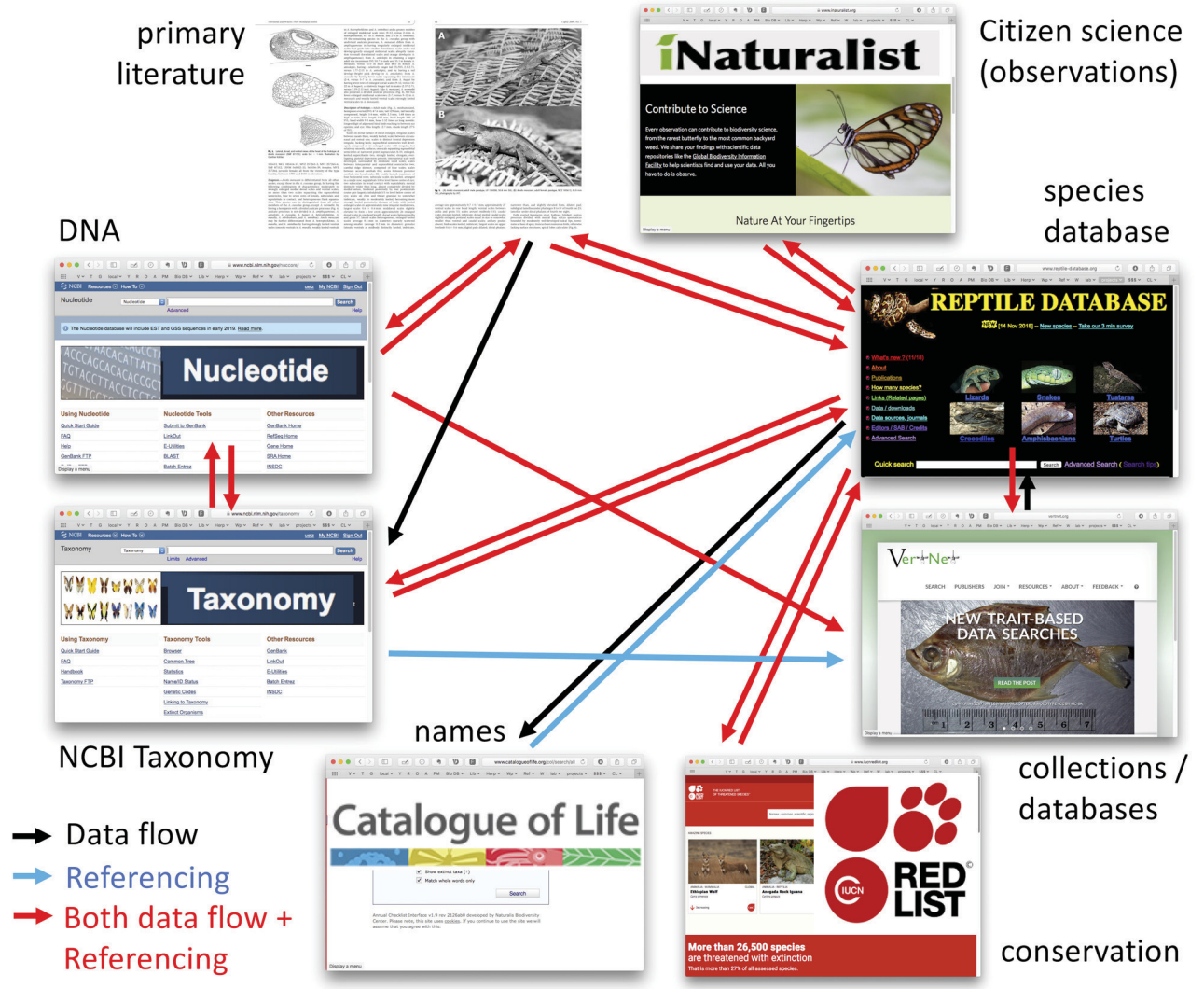

FIGURE 1. Relationship between databases, such as the NCBI Taxonomy, the Reptile Database, and other taxonomic databases and data sources. Some of the resources shown are examples for other data resources not shown. Arrows indicate data flow *or* referencing (linking) between databases. For instance, iNaturalist uses names from the Reptile Database, with the latter linking back directly to iNaturalist species pages. Arrows to and from collections (and collection data aggregators such as VertNet) indicate cross-referencing of specimens, but not necessarily hyperlinks.

The Reptile Database and GenBank's taxonomy database

The Reptile Database (http://www.reptile-database.org/) is a comprehensive taxonomy database for reptiles and provides organism names, classification, type information, taxonomic opinions, about 13,000 images and ~49,000 literature references (Uetz & Stylianou 2018). The August 2019 release of the Reptile Database recognizes >11,000 species. Information in the Reptile Database is curated directly from the literature, obtained from data submitters, or imported from other databases (including NCBI taxonomy, but also from supplementary information published with primary research papers). Taxonomic decisions are based on published information; in cases of taxonomic disagreement experts or the Scientific Advisory Board of the Reptile Database are consulted. While one of us (PU) is the main curator of the Reptile Database a team of volunteer curators helps with this effort on an irregular basis.

GenBank is a database that contains the publicly available nucleotide sequences and the names of their source organisms, among other information (Benson *et al.* 2017). Data are exchanged daily with the two other partners in the International Nucelotide Sequence Database Collaboration (INSDC) (Karsch-Mizrachi *et al*. 2017), the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) and ensures worldwide coverage. All associated organismal information is stored in the NCBI taxonomy database (Federhen 2012) and serves as a central organizing hub for INSDC as well as other resources at NCBI. It is maintained by a small group of curators and relies heavily on input from the scientific community and the availability of on-line organismal databases such as the Reptile Database for verifying submitted organism names, synonymy and classification. Unfortunately, time-consuming and error-prone manual curation is hard to avoid as long as the only source of information is the original literature, hence we argue in this paper that this step should be handled by the expert taxonomists and the authors who submit sequences to GenBank.

How many reptile names are in GenBank?

As of July 10, 2019, GenBank contained over one million sequences from 8,955 reptile "taxa" (**Fig. 2, Supplementary Table S1**). However, only about 7,400 (67%) of these have binary species names that match those in the Reptile Database (**Fig. 2**).

In addition to about 7,400 reptile names that are found in the both databases with the exact same name in both databases, another 167 differ in NCBI and the Reptile Database, being either synonyms, spelling variants, or show differences in status (e.g., species *versus* subspecies).

GenBank recognizes certain synonyms more often than the Reptile Database, because sequences have been submitted under multiple names. While homotypic (objective) synonyms (chresonyms) are based on the same type specimen, heterotypic synonyms are based on distinct type specimens. Whether these specimens belong to the same species may be contentious among researchers. For example, *Brookesia antoetrae* Brygoo & Domergue, 1971 and *B. thieli* Brygoo & Domergue, 1971 are based on different type specimens and treated as two different species in the GenBank taxonomy database but they are treated as synonyms in the Reptile Database. In this case, individual authors differ in their assessment of synonymy (some, but not all authors, consider *Brookesia antoetrae* as a synonym of *Brookesia thieli*). As of July 2019, 112 (1.2%) of the GenBank species names are listed as synonyms in the Reptile Database (**Table 1**).

TABLE 1. GenBank taxa not assigned to currently recognized species in the Reptile Database and a rough classification into different types.

NCBI taxon type	Example	Tax ID	2016	2018	2019
aff. + cf.	*Paracontias* aff. *tsararano*	665569	148	190	219
hybrid	*Thamnophis butleri* x *radix*	925770	22	27	29
sp. (incl. BOLD)	*Pelomedusa sp. A CK-2014*	1510148	1,047	1,166	1,317
synonym	*Cordylus tasmani*	884335	218	99	112
Sum			**1,435**	**1,491**	**1,677**

The "aff." and "cf." designation indicate that a species is similar but not identical to the species named in the epithet. Table 1, continued. The acronym "sp." is almost exclusively used for unknown species or temporary names (which, nevertheless, rarely get updated). BOLD = Barcode of Life Database (Adamowicz 2015; Chambers & Hebert 2016). Synonyms are mostly heterotypic as homotypic synonyms and misspellings have been cleaned up prior to publication. Data was downloaded from NCBI in May 2016, July 2018, and July 2019, respectively. A complete list of all 8,955 NCBI reptile names and their status in the Reptile Database (as of July 10, 2019) is provided in **Supplementary Table S1**.

GenBank lists another 219 species with the genus name followed by an indicator of uncertainty when a species is not unequivocally identified, such as "cf." or "aff.", e.g. *Geckoella* cf. *deccanensis* (**Table 1**).

The largest group of informal reptile names at GenBank are 1,317 'sp.' Names, e.g. *Pelomedusa* sp. C CK-2014 (**Table 1**). These provisional names come from species that have not been identified to the species level at the time of sequence submission or they come from diagnosed animals whose names have not formally been published. A subset of these names, those with a recognizable intent of the submitter to formally publish a new species name in the future is tagged as an "unpublished name" in the NCBI taxonomy DB (Schoch *et al.* 2017). Once a month, NCBI runs a semi-automatic scan of the literature for these types of names and updates the provisional name in the sequence entries with the newly published formal species name. However, this method is not comprehensive and easily misses updated names, particularly if the name submitted to the database was amended during the publication process. In the past two years, GenBank has been able to update about 80 provisional reptile species names through its journal scanning operation while there are another 90 reptile species names that are tagged as unpublished but can currently not be connected with validly published binary names.

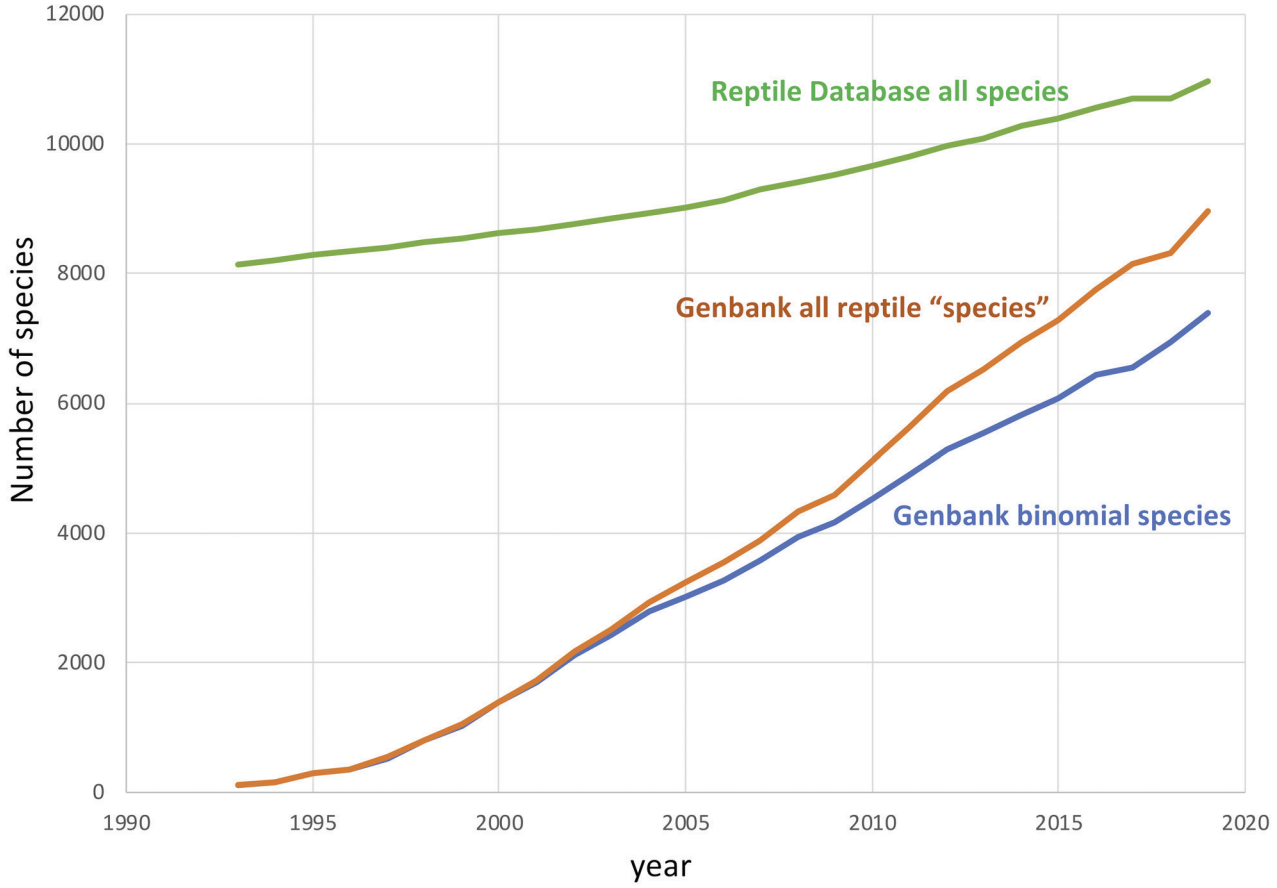

FIGURE 2. Growth of reptile species numbers in the Reptile Database and NCBI Taxonomy/ GenBank. "Genbank binomial species" are those that have a valid equivalent in the Reptile Database while GenBank all reptile "species" include all other unique names, even when their identity is not clear (or not reported to GenBank, see **Table 1**).

Sequence from type

A species name associated with a DNA sequence does not necessarily mean that the source organism was correctly identified. In fact, misidentifications happen all the time – the synonymy and comment sections in the RDB has documented numerous such cases. For instance, reports of *Boiga cynodon* (Boie, 1827) from Cambodia, Laos, and Vietnam are often based on misidentified *B. ocellata* Kroon, 1973 (Tillack *et al*. 2004), with the latter being a synonym of *B. siamensis* Nutaphand, 1971. The most important tools for reliable species identification are sequences from type material and these sequences can be retrieved from GenBank with the *sequence from type* filter in NCBI's data retrieval system ENTREZ (Federhen 2015).

As of July 2019, the NCBI query (https://www.ncbi.nlm.nih.gov/search/) "reptiles [orgn] AND sequence from type [filter]" retrieves 1757 sequence records from 365 reptile species that are from holotype or paratype material. That is just about 6% of the 5,623 reptile species which have type material information stored in GenBank's taxonomy database, but for the remaining species, there is either no sequence from type material available or the sequence submitter did not provide that information.

For example, accession MF154856 contains the RAG1 sequence from the holotype of the gecko *Goggia matzikamaensis* Heinicke, Turk & Bauer, 2017, annotated in the GenBank record with the specimen_voucher and the type_material qualifier:

/specimen_voucher="MCZ:R-192186"
/type_material="holotype of *Goggia matzikamaensis*"

where MCZ:R is the museum and collection abbreviation (Museum of Comparative Zoology, Harvard University, Reptile collection), followed by the specimen number.

While only very little type information in GenBank has come from submitter input or from the primary literature, type information in GenBank is cross-checked with type data in the Reptile Database. Sequences from type are often not identified at the time of sequence submission and that information is therefore missing from the GenBank sequence records. In these cases, the database searches demonstrated above will fail even if the type material identifier is in the taxonomy database because the necessary matching identifier in the GenBank record is not present. GenBank submitters are therefore encouraged to identify sequences from type material when submitting their sequences to GenBank by using the */note* option and use GenBank's BioCollection database (https://www.ncbi.nlm.nih.gov/biocollections) (Sharma *et al*. 2018) to check for the correct collection name.

Recently, NCBI has begun to verify taxonomic identities in prokaryotes based on average sequence identity of type strain genomes (Ciufo *et al*. 2018). Future work will extend this project to eukaryotes but it will likely be many years before type genome based comparisons can help correct mis-assignments in reptiles or animals in general.

Conclusions and recommendations

With increasing species splitting and more fine-grained population analyses, it becomes increasingly important that submitters of DNA sequences provide as much information as possible at submission time, including type information. Even more important, submitters should notify GenBank of changes that occur after sequence submission including the formal publication of new names and name changes because of taxonomic revisions or misidentifications. Readers are referred to the tools listed under the Submissions tab in https://www.ncbi.nlm.nih.gov/guide/all/.

While we have made an effort to standardize reptile names in both the Reptile Database and the reptile section of the NCBI taxonomy since 2017, much work remains to be done. Apart from the fact that the NCBI Taxonomy group cannot curate a large body of taxonomic literature themselves, experts and data submitters are much more familiar with both the data and the literature and thus most qualified for this task.

GenBank encourages submitters and co-authors to provide updates and revisions, including citation information, via email (gb-admin@ncbi.nlm.nih.gov) when formal names are published. GenBank Taxonomy curators following the latest literature and expert opinion will assign a currently accepted name and classification which will be propagated to all records in GenBank containing that name. For updates from third parties, including erroneously labeled or misidentified records, GenBank staff will contact the original submitter to verify the modification. If the submitter cannot be contacted, the GenBank staff in consultation with Taxonomy curators may flag the record as UNVERIFIED Organism, which includes removal from BLAST databases or remove it from public view as a contaminant (Benson *et al*. 2012).

NCBI and taxonomic databases need to work with collections and other stakeholders to update and continuously cross-reference their databases. While databases are working to establish robust mechanisms for data exchange it will be increasingly difficult to keep track of taxonomic and biological data without the help of the original data providers. No matter at which entry point, the original data needs to be provided by the researchers who produced the data. This could be accomplished by a taxonomic database or by NCBI but since NCBI is much more centralized than the myriad of taxonomic databases, it is probably the best solution at this time. As a minimum, we ask authors

to revisit their publications and GenBank submission for sample IDs without proper species identifications (such as "*Liolaemus* sp. 4 PW-2018 = TaxonID 2488882). Often, such manuscript names get assigned to proper species in subsequent studies without ever being updated in GenBank or the NCBI taxonomy database – but often it is only the authors who are aware of such updates. Last but not least, authors and submitters should have a vested interest in their data being properly represented in the literature and the public databases curated from it. Clarification and updating public data will tremendously help future investigations but also downstream activities such as political decisions, e.g. in conservation.

Acknowledgments

DDL thanks Stacy Ciufo, Kathleen O'Neill and Conrad Schoch for help with Entrez searches, wrangling the output and helpful suggestions for the write-up. Virginia Commonwealth University (VCU) is acknowledged for their support of the Reptile Database and PU. AG was a graduate student in Bioinformatics at VCU. The work of Detlef Leipe was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, National Institutes of Health.

References

Adamowicz, S.J. (2015) International Barcode of Life: evolution of a global research community. *Genome*, 58, 151–162. https://doi.org/10.1139/gen-2015-0094

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2017) GenBank. *Nucleic Acids Research*, 45, D37–D42. https://doi.org/10.1093/nar/gkw1070

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J. & Sayers, E.W. (2012) GenBank. *Nucleic Acids Research*, 40, D48–D53. [database issue] https://doi.org/10.1093/nar/gkr1178

Chambers, E.A. & Hebert, P.D. (2016) Assessing DNA barcodes for species identification in North American reptiles and amphibians in natural history collections. *PLoS One*, 11, e0154363. https://doi.org/10.1371/journal.pone.0154363

Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., Brover, S., Schoch, C.L., Kimchi, A. & DiCuccio, M. (2018) Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *International Journal of Systematic and Evolutionary Microbiology*, 68, 2386–2392. https://doi.org/10.1099/ijsem.0.002809

Federhen, S. (2012) The NCBI Taxonomy database. *Nucleic Acids Research*, 40, D136–143. https://doi.org/10.1093/nar/gkr1178

Federhen, S. (2015) Type material in the NCBI taxonomy database. *Nucleic Acids Research*, 43, D1086–98. https://doi.org/10.1093/nar/gku1127

Heinicke, M.P., Turk, D. & Bauer, A.M. (2017) Molecular phylogeny reveals strong biogeographic signal and two new species in a Cape Biodiversity Hotspot endemic mini-radiation, the pygmy geckos (Gekkonidae: Goggia). *Zootaxa*, 4312 (3), 449–470. https://doi.org/10.11646/zootaxa.4312.3.3

Karsch-Mizrachi, I., Takagi, T. & Cochrane, G. (2017) The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 46 (D1), D48–D51. https://doi.org/10.1093/nar/gkx1097

Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D. & Karsch-Mizrachi, I. (2019) GenBank. *Nucleic Acids Research*, 47 (D1), D94–D99. https://doi.org/10.1093/nar/gky989

Schoch, C.L., Aime, M.C., de Beer, W., Crous, P.W., Hyde, K.D., Penev, L., Seifert, K.A., Stadler, M., Zhang, N. & Miller, A.N. (2017) Using standard keywords in publications to facilitate updates of new fungal taxonomic names. *IMA Fungus*, 8 (2), 70–73. https://doi.org/10.1007/BF03449466

Sharma, S., Ciufo, S., Starchenko, E., Darji, D., Chlumsky, L., Karsch-Mizrachi, I. & Schoch, C.L. (2018) The NCBI BioCollections Database. *Database, Oxford*, 2018, bay006. https://doi.org/10.1093/database/bay006

Tillack, F., Ziegler, T. & Le Khac Quyet (2004) Eine neue Art der Gattung *Boiga* Fitzinger 1826 (Serpentes: Colubridae: Colubrinae) aus dem zentralen Vietnam. *Sauria*, 26 (4), 3–13.

Uetz, P. & Garg, A. (2017) Molecular taxonomy: Species disconnected from DNA sequences. *Nature*, 545 (7655), 412. https://doi.org/10.1038/545412c
Uetz, P., Freed, P. & Hošek, J. (2019) The Reptile Database. Available from: http://www.reptile-database.org (accessed 10 July 2019)
Uetz, P. & Stylianou, A. (2018) The original descriptions of reptiles and their subspecies. *Zootaxa*, 4375 (2), 257–264. https://doi.org/10.11646/zootaxa.4375.2.5

SUPPLEMENTARY TABLE S1. All reptile names from the NCBI Taxonomy Database as of May 2016, July 2018, and July 2019 with their corresponding names in the Reptile Database. Available at http://www.reptile-database.org/data/Garg2019_Table_S1.xlsx.